結論部の含意を考慮したルール抽出法 (数値最適化の理論と実際)

AUTHOR(S):

楠木, 祥文; 井上, 正則; 乾口, 雅弘

CITATION:

楠木, 祥文 ...[et al]. 結論部の含意を考慮したルール抽出法 (数値最適化の理論と実際). 数理解析研究所講究録 2008, 1584: 59-71

ISSUE DATE:

2008-02

URL:

http://hdl.handle.net/2433/81500

数理解析研究所講究録
第1584巻 2008年 59-71

結論部の含意を考慮したルール抽出法

大阪大学大学院基礎工学研究科 楠木 祥文 (Yoshifumi Kusunoki)
Graduate School of Engineering Science, Osaka University
大阪大学基礎工学部 井上 正則 (Masanori Inoue)
School of Engineering Science, Osaka University
大阪大学大学院基礎工学研究科 乾口 雅弘 (Masahiro Inuiguchi)
Graduate School of Engineering Science, Osaka University

1. はじめに

ラフ集合理論 [8, 9] は識別不能性という不確実性を扱う数学的アプローチであり，データ解析のための様々なツールが提案されている [7]．ラフ集合理論では，データは決定表で表わされる．決定表はいくつかの属性によって特徴づけられた対象の集合で構成されており，各対象は一つの決定クラスに分類されている．ラフ集合理論で提案されているツールの一つに決定表からのルール抽出 [3, 5] があり，抽出されたルールは，どのような属性値をもつ対象がどのクラスに帰属するかを推論する．本研究ではルール抽出手法を取り扱う．

与えられた決定表の決定クラスに順序があり，いくつかの属性の値にも順序があるとする．さらに，最も高い決定クラスの周りに次に高いクラスが存在し，その周りにその次に高いクラスが存在していくような入れ子構造があると仮定する．たとえば，料理における調味料の量と好みに関する決定表はこの仮定にあてはまる．つまり，ほどよい調味料の量が好ましく，その量から離れるほど好まれなくなる．このような決定表からは決定クラス Cl_i の帰属を直接推定するルールを抽出するより，Cl_i 以上や Cl_i 以下の帰属を推定するルールを抽出した方が好ましいと考えられる．少なくともより簡潔なルールとなる．Cl_i 以上や Cl_i 以下の帰属を推論するルールを考えると，結論部に含意関係が成立すれば条件部にも含意関係が成立した方が望ましい．つまり，三つの決定クラスがあり，$Cl_3 \succ Cl_2 \succ Cl_1$ という順序付けがなされている場合，Cl_3 以上であると推論される対象は Cl_2 以上であると推論されるべきである．しかし，従来のルール抽出法では独立にルールが抽出されるため，この関係が必ず成り立つとは限らない．そこで，本研究では結論部の含意関係を考慮したルール抽出法を三つ提案する．さらに，抽出されたルールを用いた未知対象の分類法も提案する．最後に数値実験を行い，提案法の有用性を分類精度などによって評価する．

2. ラフ集合とルール抽出

2.1. ラフ集合と決定表

ラフ集合理論に基づきデータから決定ルールを抽出する問題では，データは決定表の形式で与えられる．決定表は4項対 $\langle U, C \cup \{d\}, V, \rho \rangle$ で表現され，U は対象の有限集合，C は条件属性の有限集合，d は決定属性，V は属性値の集合，$\rho : U \times C \cup \{d\} \to \bigcup_{a \in C \cup \{d\}} V_a$ は情報関数である．ここで，V_a は属性 a が取りうる属性値の集合である．また，決定属性 d の値が同じ対象の集合を決定クラスと呼び，Cl_i で表わす．対象の集合 U は決定クラス Cl_i により分割される．すなわち，$\boldsymbol{Cl} = \{Cl_1, Cl_2, ..., Cl_p\}$ は U の分割になっている．表1に決

定表の例をあげる．表1は自動車の評価に関するデータであり，各行には，その行に対応する対象の属性値が割り当てられている．たとえば対象u_1の属性値は，$\rho(u_1, \text{価格}) = 500$, $\rho(u_1, \text{維持費}) = 75$, $\rho(u_1, \text{安全性}) = \text{高}$，そして$\rho(u_1, \text{評価}) = \text{悪い}$である．この表の決定クラスは，$Cl_1 = \{x \in U \mid \rho(x,d) = \text{悪い}\}$, $Cl_2 = \{x \in U \mid \rho(x,d) = \text{普通}\}$, $Cl_3 = \{x \in U \mid \rho(x,d) = \text{良い}\}$である．

表 1: 自動車の評価に関する決定表

	条件属性			決定属性
対象	価格 (万円)	維持費 (万円)	安全性	評価
u_1	500	75	高	悪い
u_2	300	61	低	悪い
u_3	230	51	中	悪い
u_4	230	51	中	普通
u_5	210	60	低	普通
u_6	180	62	高	普通
u_7	120	45	高	良い
u_8	80	50	中	良い
u_9	60	44	中	良い

属性の部分集合$A \subseteq C \cup \{d\}$が与えられると，Aの属性値がまったく同じ対象の対$(x,y), x,y \in U$は識別できない．ラフ集合理論はこの識別不能性に基づいてデータを解析する．データの解析を行うために，まず対象間の識別不能関係$R_A \subseteq U \times U$を次のように定義する．

$$R_A = \{(x,y) \mid \rho(x,a) = \rho(y,a), \forall a \in A\} \tag{1}$$

R_Aは，反射性，推移性，対称性を満たすので同値関係である．同値関係R_Aから同値類を次のように定義できる．

$$[x]_A = \{y \in U \mid (y,x) \in R_A\} \tag{2}$$

本研究の目的は条件属性値から決定クラスを推論する決定ルールを求めることである．与えられた条件属性を用いて決定クラスを正確に特徴づけられればよいが，すべての条件属性を用いても特徴づけられない決定クラスが存在する場合がある．その場合，これらの決定クラスを完全に表現する無矛盾な決定ルール群を求めることはできない．ラフ集合理論では，条件属性により定義できる上近似と下近似という二つの集合を用いる．条件属性集合Cに対して，決定クラスCl_iの上近似$C^*(Cl_i)$と下近似$C_*(Cl_i)$は次のように定義される．

$$C^*(Cl_i) = \{x \in U \mid [x]_C \cap Cl_i \neq \emptyset\} \tag{3}$$

$$C_*(Cl_i) = \{x \in U \mid [x]_C \subseteq Cl_i\} \tag{4}$$

上近似を用いるとその決定クラスに帰属しうる対象を示すルール，下近似を用いると決定クラスに確実に帰属する対象を示すルールが得られる [4]．表1の評価が悪いクラスCl_1は，u_3とu_4が矛盾しているので，条件属性をすべて用いても表現することはできないが，その上近似$C^*(Cl_1) = \{u_1, u_2, u_3, u_4\}$と，下近似は$C_*(Cl_1) = \{u_1, u_2\}$に関するルールは，たとえば，次のように求めることができる．

If $\rho(x, \text{維持費}) \geq 51$ **and** $\rho(x, \text{安全性}) =$ 中 **then** x は Cl_1 に帰属しうる

If $\rho(x, \text{価格}) \geq 300$ **then** x は Cl_1 に確実に帰属する

本研究では，決定クラスに関して，$Cl_p \succ Cl_{p-1} \succ ... \succ Cl_1$なる順序関係があると仮定する．このとき，上側和集合と下側和集合はそれぞれ，

$$Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s, \ t = 1, 2, ..., p \tag{5}$$

$$Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s, \ t = 1, 2, ..., p \tag{6}$$

と定義される．これらの和集合の上近似または下近似から決定ルールを抽出することを考える．これにより，たとえば，表1から次のようなルールが抽出できる．

If $\rho(x, \text{価格}) \geq 230$ **then** x は $Cl_2^{\geq}$ に確実に帰属する

2.2. MLEM2

本研究では決定ルールを抽出するアルゴリズムとして MLEM2[6] を用いる．MLEM2 はラフ集合理論に基づくデータマイニングシステム LERS のサブシステムであり，各決定クラスの下近似データを入力すると確実性決定ルール，上近似データを入力すると可能性決定ルールを抽出することができる．数値属性をもつ決定表から決定ルールを抽出する場合，数値属性を離散化する必要がある [5] が，MLEM2 はルール抽出と並行して離散化も行う．

MLEM2 アルゴリズムを説明する前にいくつか記号の説明をする．決定ルールの条件を3項対$t = (a, v, R)$で表す．ここで，aは条件属性，vは条件属性値，Rは関係である．aが名義属性ならばRは$=$，数値属性や順序を持つ属性ならばRは$\geq$または$<$とする．条件tに対し，tを満たす対象の集合を$[t] = \{x \in U \mid \rho(x, a) \ R \ v\}$と表す．決定ルールの条件部はいくつかの条件の連言をとったものであるので，条件の集合を条件部とみなすことでき，さらに，結論部が特定されている状況では決定ルールとみなすことができる．決定ルールTに対して，Tを満たす対象の集合を$[T] = \bigcap_{t \in T}[t]$と表す．

MLEM2 のアルゴリズムを以下に示す．まず，決定ルールの集合$\mathbb{T}$を初期化して，$\mathbb{T}$に含まれるどのルールにもカバーされていない対象集合Gに入力Bを代入する．ここで，決定ルールTが対象xをカバーするとは，$x \in [T]$のことを示す．**while** $G \neq \emptyset$ のループでは，すべての入力対象が$\mathbb{T}$のどれかの決定ルールにカバーされるまでルールTを$\mathbb{T}$に追加し続ける．このループの中に入ると，まず，決定ルールTを初期化し，ルールTの要素になりうる条件の集合$T(G)$を求める．**while** $T = \emptyset$ or $[T] \not\subseteq B$ のループでは，決

定ルールがカバーする集合に入力B以外の対象が含まれなくなるまで条件tをある基準にしたがって選択しTに追加し続ける．決定ルールTを生成した後，**for** each t in TのループでTから冗長な条件tを取り除き，Tを$\mathbb{T}$に加える．そして，まだカバーされていない対象集合Gを更新する．**while** $G \neq \emptyset$のループを抜けた後，**for** each T in $\mathbb{T}$のループで$\mathbb{T}$から冗長な決定ルールTを取り除き，$\mathbb{T}$を出力する．

Procedure MLEM2
(**input:** B,
output: $\mathbb{T}$)
begin
  $\mathbb{T} := \emptyset$;
  $G := B$;
  while $G \neq \emptyset$ **do begin**
    $T := \emptyset$; $T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
    while $T = \emptyset$ **or** $[T] \not\subseteq B$ **begin**
      † select $t \in T(G)$ with the highest priority, if a tie occurs,
      select $t \in T(G)$ such that $|[t] \cap G|$ is maximum; if another tie occurs,
      select $t \in T(G)$ with smallest cardinality of $[t]$; if a further tie occurs,
      select a first one;
      $T := T \cup \{t\}$;
      $G := [t] \cap G$;
      $T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
      $T(G) := T(G) - T$;
    end {while};
    for each t in T **do begin**
      if $[T - \{t\}] \subseteq B$ **then** $T := T - \{t\}$;
    end {for};
    $\mathbb{T} := \mathbb{T} \cup \{T\}$;
    $G := B - \bigcup_{T \in \mathbb{T}} [T]$;
  end {while};
  for each T in $\mathbb{T}$ **do begin**
    if $\bigcup_{S \in \mathbb{T} - \{T\}} [S] = B$ **then** $\mathbb{T} := \mathbb{T} - \{T\}$;
  end {for};
end {Procedure};

2.3. 未知データの分類

MLEM2により決定ルールが抽出されると，これに基づき決定属性値が未知の対象が帰属する決定クラスを推定することが考えられる．未知対象の条件属性データが満たす条件をもつ決定ルールのすべてが共通に一つのクラスへの帰属を示せば良いが，異なった決定クラスへの帰属を示す場合がある．一方，未知対象の条件属性データがいずれの決定ルールの条件も満たさない場合もある．これらの場合に対処する方法がLERS[5, 6]の中で与えられている．

未知対象の条件属性値が少なくとも一つの決定ルールの条件を満たす場合には，*strength*, *specificity*の二つに基づく評価基準*support*によっていずれの決定クラスに帰属するかを定める．$strength(r)$はルールrの例となる学習用データの数，$specificity(r)$はルールrに含まれる条件の数(条件の長さ)である．このとき，$support_{Cl}(x)$は未知対象xの条件

属性データが満足し，かつ，結論部が決定クラスClへの帰属を示す決定ルールに関して$strength$と$specificity$の積を合計した値であり，次式で定められる．

$$support_{Cl}(x) = \sum_{matching\ rules\ r\ inferring\ Cl} strength(r) * specificity(r) \tag{7}$$

この$support_{Cl}(x)$が最も大きい決定クラスClへ対象が分類される．

未知対象の条件属性値がいずれの決定ルールの条件属性を満たさない場合には，一部の条件が満たされるルール(部分合致ルール)が用いられる．部分合致ルールrに対して，ルールrの条件部に含まれる条件の数に対する未知対象が満たす条件の割合である$matching_factor(r)$が算出され，次式の値が最も大きい決定クラスに未知対象が分類される．

$$p\_support_{Cl}(x) = \sum_{\substack{partially\ matching\ rules\ r \\ inferring\ Cl}} \begin{array}{l} matching\_factor(r) \\ * strength(r) * specificity(r) \end{array} \tag{8}$$

3. 結論部の含意関係を考慮したルール抽出法

決定ルールの結論部の含意関係を考慮したMLEM2ベースのルール抽出法を三つ提案し，各小節で詳しく説明する．上側和集合$Cl_s^{\geq}$, $Cl_t^{\geq}$について，$x \in Cl_s^{\geq} \Rightarrow x \in Cl_t^{\geq}$が成立する場合，はじめに述べる二つの手法を用いると$Cl_t^{\geq}$から得らたルール群は$Cl_s^{\geq}$から得られたルール群を包含する．ここで，$RULES_1$, $RULES_2$をそれぞれルール群つまり決定ルールの集合とすると，$RULES_1$が$RULES_2$を包含するとは，$RULES_2$に含まれる一つ以上の決定ルールにカバーされる条件属性パターン$c \in \prod_{a \in C} V_a$は，$RULES_1$に含まれる一つ以上の決定ルールにカバーされることを表す．三つ目の手法ではその保証はないがデータの構造を考慮したルール抽出が行われる．

3.1. 制限法

制限法は無矛盾な目標の対象集合Bとルール群$\mathbb{T}_0$を入力とし，$\mathbb{T}_0$に包含され，かつBをカバーするルール群$\mathbb{T}$を出力する．ただし，Bは$\mathbb{T}_0$にカバーされている必要がある．アルゴリズムを以下に示す．このアルゴリズムにはMLEM2と異なる部分が二つある．一つは，探索する条件部Tの初期化の部分である．MLEM2では$T = \emptyset$としたが，制限法では$T_0 \in \mathbb{T}_0$をTの初期値とする．初期値T_0の選択は‡で行われ，T_0は三つの基準を辞書式に適用することで評価される．これは条件tを$T(G)$から選択する場合と同じである．第一基準では，$\mathbb{T}$にカバーされていない対象集合Gの中でT_0がカバーしている対象の数で大きいものが優先される．第二基準では，T_0がカバーしている対象の数で小さいものが優先される．第三基準では，はじめに選択されたものが優先される．もう一つは，探索された条件部Tから冗長な条件を取り除く部分である．制限法では，TからT_0の条件を取り除いた後，Tから冗長な条件を取り除く．そして，for each t in T_0のループでTがT_0に包含されるようにTに条件tを加える．$t = (a, v, R) \in T_0$に対して$\hat{t} = \{(a, w, R) \mid w \in V_a\}$であり，$\hat{t} \cap T = \emptyset$の場合，$t$を$T$に加えることにより$T$が$T_0$に包含されるようになる．

Procedure MLEM2_Restriction
(**input:** B, $\mathbb{T}_0$
output: $\mathbb{T}$)
begin
  $\mathbb{T} := \emptyset$;
  $G := B$;
  while $G \neq \emptyset$ **do begin**
    ‡ $T_0 \in \mathbb{T}_0$ such that $|[T_0] \cap G|$ is maximum; if another tie occurs,
      select $T_0 \in \mathbb{T}_0$ with smallest cardinality of $[T_0]$; if a further tie occurs,
      select a first one;
    $T := T_0$;
    $G := [T] \cap G$;
    $T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
    $T(G) := T(G) - T$;
    while $[T] \not\subseteq B$ **begin**
      † select $t \in T(G)$ such that $|[t] \cap G|$ is maximum; if another tie occurs,
        select $t \in T(G)$ with smallest cardinality of $[t]$; if a further tie occurs,
        select a first one;
      $T := T \cup \{t\}$;
      $G := [t] \cap G$;
      $T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
      $T(G) := T(G) - T$;
    end {while};
    $T := T - T_0$;
    for each t in T **do begin**
      if $[(T - \{t\}) \cup T_0] \subseteq B$ **then** $T := T - \{t\}$;
    end {for};
    for each t in T_0 **do begin**
      if $\hat{t} \cap T = \emptyset$ **then** $T := T \cup \{t\}$;
    end {for};
    $\mathbb{T} := \mathbb{T} \cup \{T\}$;
    $G := B - \bigcup_{T \in \mathbb{T}} [T]$;
  end {while};
  for each T in $\mathbb{T}$ **do begin**
    if $\bigcup_{S \in \mathbb{T} - \{T\}} [S] = B$ **then** $\mathbb{T} := \mathbb{T} - \{T\}$;
  end {for};
end {Procedure};

3.2. 緩和法

緩和法は無矛盾な目標の対象集合 B とルール群 $\mathbb{T}_0$ を入力とし，$\mathbb{T}_0$ を包含し，かつ B をカバーするルール群 $\mathbb{T}$ を出力する．ただし，$\mathbb{T}_0$ は $U - B$ をカバーしていないものに限る．アルゴリズムを以下に示す．緩和法は大きく分けて $\mathbb{T}_0$ を包含するルール群 $\mathbb{T}$ を求める部分と $\mathbb{T}$ がカバーできなかった対象集合をカバーするルール群 $\mathbb{T}'$ を求める部分から構成される．はじめの **for** each T_0 in $\mathbb{T}_0$ のループで，$\mathbb{T}_0$ の各決定ルールを包含する決定ルールを求める．MLEM2 と異なる点は $T(G)$ の構成方法であり，$t \in T(G)$ は $\{t\}$ が T_0 を包含するものに限定されている．これにより，ループの中で生成される T は T_0 を包含するものになる．**for** each T_0 in $\mathbb{T}_0$ のループを抜けたあと，$\mathbb{T}$ の中で冗長な決定ルールつ

まり他のルールに包含されているものを取り除き，$\mathbb{T}$ではカバーされていない対象集合Gを求める．Gをカバーするルール群$\mathbb{T}'$をMLEM2で求め，それを$\mathbb{T}$に加え，出力する．

Procedure MLEM2_Relaxation
(**input:** B, $\mathbb{T}_0$
output: $\mathbb{T}$)
begin
  for each T_0 in $\mathbb{T}_0$ **do begin**
    $\mathbb{T} := \emptyset$;
    $G := B$;
    $T := \emptyset$;
    $T(G) := \{t \mid [t] \cap G \neq \emptyset,\ \{t\}$ is includes $T_0\}$;
    while $T = \emptyset$ **or** $[T] \not\subseteq B$ **begin**
      † select $t \in T(G)$ such that $|[t] \cap G|$ is maximum; if another tie occurs, select $t \in T(G)$ with smallest cardinality of $[t]$; if a further tie occurs, select a first one;
      $T := T \cup \{t\}$;
      $G := [t] \cap G$;
      $T(G) := \{t \mid [t] \cap G \neq \emptyset,\ \{t\}$ is includes $T_0\}$;
      $T(G) := T(G) - T$;
    end {while};
    for each t in T **do begin**
      if $[T - \{t\}] \subseteq B$ **then** $T := T - \{t\}$;
    end {for};
    $\mathbb{T} := \mathbb{T} \cup \{T\}$;
  end {for};
  for each T_1 in $\mathbb{T}$ **do begin**
    for each T_2 in $\mathbb{T}$ **do begin**
      if $T_1 \neq T_2$ **and** T_1 is included in T_2 **then** $\mathbb{T} := \mathbb{T} - \{T_1\}$;
    end {for};
  end {for};
  $G := B - \bigcup_{T \in \mathbb{T}} [T]$;
  if $G \neq \emptyset$ **then** MLEM2(G, $\mathbb{T}'$);
  $\mathbb{T} := \mathbb{T} \cup \mathbb{T}'$;
end {Procedure};

3.3. 優先順位法

本研究では，決定属性値に順序が存在し，決定属性値の高い値のデータの周りにより低い値のデータが存在し，その周りにさらに低いデータが存在するという入れ子構造，あるいは，逆に，決定属性の低い値のデータの周りにより高い値のデータが存在し，その周りにさらに高いデータが存在するという入れ子構造を仮定している．データがこのような構造をもつ場合，峰または谷の頂点をとらえ，それを囲むように決定ルールを抽出すれば，上位または下位のルール群を包含するようなルール群が得られると考えられる．優先順位法では，この考えに基づき，上側和集合の決定ルールを求める場合はより高い決定クラスを含むようにルールの条件を選択し，下側和集合の決定ルールを求める場合はより低い決定クラスを含むようにルールの条件を選択するようにMLEM2の条件選択基準を変更する．つまり，目標の和集合に包含される和集合の中でより小さな和集合を優先して決定ルールがカバーするように条件を選択する．優先順位法のアルゴリズムを

以下に示す．$Cl_i^{\geq}$ のルール群を求める場合を考えると，† の部分で，まず $|[t] \cap Cl_p^{\geq} \cap G|$ が最も大きい条件 t を候補とする．もし候補が複数存在すれば，次は $|[t] \cap Cl_{p-1}^{\geq} \cap G|$ が基準となり候補を選択する．これを繰り返し，候補がただ一つなればその条件を T に加える．もし，$|[t] \cap Cl_i^{\geq} \cap G|$ でも候補が一つに定まらなければ，$[t]$ の基数が最小の条件 t を候補とする．それでも候補が複数存在すれば，はじめに選択した条件を T に加える．

Procedure MLEM2_Priority
(**input:** $B = C_*(Cl^{\geq})$, $C_*(Cl^{\leq})$, $C^*(Cl^{\geq})$ or $C^*(Cl^{\leq})$,
output: $\mathbb{T}$)
begin
  $\mathbb{T} := \emptyset$;
  $G := B$;
  while $G \neq \emptyset$ **do begin**
    $T := \emptyset$;
    $T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
    while $T = \emptyset$ **or** $[T] \not\subseteq B$ **begin**
      † select $t \in T(G)$ so that t covers smaller union of decision classes preferentially; if ties occur, select $t \in T(G)$ with smallest cardinality of $[t]$; if a further tie occurs, select a first one;
      $T := T \cup \{t\}$;
      $G := [t] \cap G$;
      $T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
      $T(G) := T(G) - T$;
    end {while};
    for each t in T **do begin**
      if $[T - \{t\}] \subseteq B$ **then** $T := T - \{t\}$;
    end {for};
    $\mathbb{T} := \mathbb{T} \cup \{T\}$;
    $G := B - \bigcup_{T \in \mathbb{T}} [T]$;
  end {while};
  for each T in $\mathbb{T}$ **do begin**
    if $\bigcup_{S \in \mathbb{T} - \{T\}} [S] = B$ **then** $\mathbb{T} := \mathbb{T} - \{T\}$;
  end {for};
end {Procedure};

4. 未知データの分類手法

上側和集合と下側和集合からルール群が抽出されれば，それらを用いて未知の対象の帰属する決定クラスを推定することができる．もしルール群を用いて未知対象 x を $Cl_i^{\geq}$ かつ $Cl_i^{\leq}$ と推論できれば，x を Cl_i に分類することができる．しかし必ずしも一つの決定クラスに未知対象を分類できるとは限らず，決定クラスを特定できない場合や推論が矛盾する場合がある．たとえば，$1 \leq s < t \leq p$ のとき，x が $Cl_s^{\geq}$ かつ $Cl_t^{\leq}$ と推論されたならば，x の決定クラスを特定することはできず，また，x が $Cl_t^{\geq}$ かつ $Cl_s^{\leq}$ と推論されたならば，この推論は矛盾しているといえる．Błaszczyński et al. [2] は上側和集合や下側和集合を推論するルール群を用いた未知対象の分類方法を提案しているが，本研究ではこの方法とは異なる考え方に基づいた分類方法を提案する．

まず，決定クラスの各境界で和集合の評価値を用いて未知対象 x がどちらの和集合に分類されるかを決定する．たとえば，$Cl_t^{\geq}$ と $Cl_{t-1}^{\leq}$ の境界の場合は二つの評価値 $E_{Cl_t^{\geq}}(x)$ と $E_{Cl_{t-1}^{\leq}}(x)$ を比較する．

$$E_{Cl_t^{\geq}}(x) = -\sum_{s<t} support_{Cl_s^{\leq}}(x) \tag{9}$$

$$E_{Cl_{t-1}^{\leq}}(x) = -\sum_{s\geq t} support_{Cl_s^{\geq}}(x) \tag{10}$$

$support$ は第 2.3 節の式 (7) で定義されたものである．$Cl_s^{\leq} \cap Cl_t^{\geq} = \emptyset$, $s < t$ であるので，式 (9) の $\sum_{s<t} support_{Cl_s^{\leq}}(x)$ は x が $Cl_t^{\geq}$ に帰属することを否定する度合を表わしていると考えられる．同様に式 (10) の $\sum_{s\geq t} support_{Cl_s^{\geq}}(x)$ も x が $Cl_{t-1}^{\leq}$ に帰属することを否定する度合を表わしていると考えられる．$E_{Cl_t^{\geq}}(x) > E_{Cl_{t-1}^{\leq}}(x)$ であれば，x を $Cl_t^{\geq}$ に分類し，そうでなければ $Cl_{t-1}^{\leq}$ に分類する．ただし，$E_{Cl_t^{\geq}}(x) = E_{Cl_{t-1}^{\leq}}(x) = 0$ の場合は，次の評価値

$$E'_{Cl_t^{\geq}}(x) = -\sum_{s<t} p\_support_{Cl_s^{\leq}}(x) \tag{11}$$

$$E'_{Cl_{t-1}^{\leq}}(x) = -\sum_{s\geq t} p\_support_{Cl_s^{\geq}}(x) \tag{12}$$

を比較し，$E'_{Cl_t^{\geq}}(x) > E'_{Cl_{t-1}^{\leq}}(x)$ であれば，x を $Cl_t^{\geq}$ に分類し，そうでなければ $Cl_{t-1}^{\leq}$ に分類する．この分類を $i = 2, 3, ..., p$ について行う．

次に，各 Cl_i, $i = 1, 2, ..., p$ について，それを支持する和集合つまり Cl_i を包含する和集合に x が何回分類されたかを数え上げる．x が分類された，Cl_i を支持する上側和集合の集合を次のように定義する．

$$WIN_{Cl_i}^{\geq}(x) = \left\{ Cl_t^{\geq} \;\middle|\; \begin{array}{l} E_{Cl_t^{\geq}}(x) > E_{Cl_{t-1}^{\leq}}(x) \\ \text{or } E'_{Cl_t^{\geq}}(x) > E'_{Cl_{t-1}^{\leq}}(x),\ E_{Cl_t^{\geq}}(x) = E_{Cl_{t-1}^{\leq}}(x) = 0 \\ t = 2, 3, ..., i \end{array} \right\} \tag{13}$$

下側和集合についても同様に定義する．

$$WIN_{Cl_i}^{\leq}(x) = \left\{ Cl_{t-1}^{\leq} \;\middle|\; \begin{array}{l} E_{Cl_t^{\geq}}(x) \leq E_{Cl_{t-1}^{\leq}}(x),\ (E_{Cl_t^{\geq}}(x) \neq 0 \text{ or } E_{Cl_{t-1}^{\leq}}(x) \neq 0) \\ \text{or } E'_{Cl_t^{\geq}}(x) \leq E'_{Cl_{t-1}^{\leq}}(x),\ E_{Cl_t^{\geq}}(x) = E_{Cl_{t-1}^{\leq}}(x) = 0, \\ t = i+1, i+2, ..., p \end{array} \right\} \tag{14}$$

式 (13), (14) から Cl_i を支持する和集合に x が分類された回数は，

$$COUNT_{Cl_i}(x) = |WIN_{Cl_i}^{\geq}(x)| + |WIN_{Cl_i}^{\leq}(x)| \tag{15}$$

となるので，$COUNT_{Cl_i}(x)$ が最も大きい Cl_i に x を分類する．

5. 数値実験

5.1. 実験方法

提案したルール抽出法の有用性を評価するために数値実験を行う．実験方法を簡単に説明する．まず与えられた対象集合から訓練用の対象をサンプリングし，訓練用の対象集合からルールを抽出する．そして，各手法の一貫性と分類精度を求め，これを手法の評価値とする．ここで，一貫性とは，ある上側 (下側) 和集合に推論されるとき必ずそれより下位 (上位) の上側 (下側) 和集合に推論される対象の割合であり，上側あるいは下側和集合間の含意関係の正しさを表わしており，分類精度とは，得られたルール群を用いてすべての対象を分類したとき，正しく分類された対象の割合である．

実験データとして UCI データベース [1] より入手した自動車の評価に関するデータを用いる．対象数は 1728，属性数は 6，決定クラス数は 4 であり，決定属性値は very good, good, acceptable, unacceptable であり，very good$\succ$good$\succ$acceptable$\succ$unacceptable と仮定する．また，すべての条件属性値パターンが存在し，どのパターンも重複していない．つまり，このデータには矛盾が存在しない．

提案した分類法で対象を分類するために，すべての上側和集合と下側和集合からルール群を抽出するが，制限法と緩和法に関しては，はじめにどの和集合からルール群を抽出するかによって結果が異なってくる．そこで，本研究では制限法と緩和法の適用方法として次の二つについて実験を行う．

- 上側和集合では，$Cl_2^{\geq}$ から従来法を用いてルール群を抽出し，$Cl_3^{\geq}, ..., Cl_p^{\geq}$ の順に制限法を用いて一つ前のルール群に包含されるルール群を抽出する．下側和集合では，$Cl_{p-1}^{\leq}$ から従来法を用いてルール群を抽出し，$Cl_{p-2}^{\leq}, ..., Cl_1^{\leq}$ の順に制限法を用いて一つ前のルール群に包含されるルール群を抽出する．

- 上側和集合では，$Cl_p^{\geq}$ から従来法を用いてルール群を抽出し，$Cl_{p-1}^{\geq}, ..., Cl_2^{\geq}$ の順に緩和法を用いて一つ前のルール群を包含するルール群を抽出する．下側和集合では，$Cl_1^{\leq}$ から従来法を用いてルール群を抽出し，$Cl_2^{\leq}, ..., Cl_{p-1}^{\leq}$ の順に緩和法を用いて一つ前のルール群を包含するルール群を抽出する．

便宜上，前者を制限法，後者を緩和法と呼ぶ．本研究では制限法と緩和法に加え優先順位法と従来法についても実験を行う．従来法とは各上側または下側和集合から独立にルール群を求める方法であり，提案法との比較に用いる．

訓練用の対象数が小さいと，ある上側 (下側) 和集合に推論された対象がより下位 (上位) の上側 (下側) 和集合に推論されない場合が多くなり，提案法と従来法の違いが顕著に表れると考えられる．そこで，全対象集合からサンプリングする割合が変化したときの分類精度の変化を観測する．全対象集合に対するサンプリングの割合は，1%から 9%の間では 1%刻み，10%から 90%の間では 10%刻みで変化させる．各割合で，1000 回実行する．

5.2. 実験結果と考察

以下に提示される表中の値は 1000 回の実験の平均値と標準偏差を表わしている．各サンプリングの割合について，各提案法と従来法の母平均に差がないという帰無仮説を立

表 2: ルール群の一貫性

割合 (%)	優先順位法 (%)	従来法 (%)
1	91.21 ± 8.34	90.37 ± 8.73
2	91.53 ± 5.58	90.95 ± 5.78
3	92.69 ± 4.15	92.01 ± 4.30
4	93.88 ± 3.19	93.18 ± 3.37
5	94.52 ± 2.73	93.85 ± 2.79
6	95.22 ± 2.40	94.49 ± 2.57
7	95.75 ± 2.08	95.11 ± 2.18
8	96.23 ± 1.88	95.68 ± 2.07
9	96.62 ± 1.66	96.10 ± 1.77
10	96.87 ± 1.56	96.40 ± 1.63
20	98.44 ± 0.80	98.26 ± 0.84
30	98.94 ± 0.52	98.86 ± 0.57
40	99.25 ± 0.40	99.20 ± 0.41
50	99.44 ± 0.32	99.40 ± 0.34
60	99.55 ± 0.27	99.51 ± 0.27
70	99.64 ± 0.23	99.62 ± 0.22
80	99.69 ± 0.19	99.68 ± 0.20
90	99.75 ± 0.16	99.73 ± 0.17

て，有意水準 5%で対応のある t-検定を行い，帰無仮説が棄却されない場合は，その検定に対応する表の値に * を付加している．

まず，表 2 に抽出されたルール群の一貫性を示す．制限法と緩和法の一貫性は 100%であり，表 2 では優先順位法と従来法のみを示す．優先順位法と従来法を比較すると，すべてのサンプリングの割合について，有意水準 5%の t-検定で有意差があり，優先順位法は従来法よりも一貫性が大きい．

次に，分類精度と各提案法と従来法の分類精度の差を表 3 と図 1 にそれぞれ示す．

制限法と緩和法について見ると，サンプリングの割合が 10%より小さい場合では，従来法より分類精度が低下している．制限法と緩和法は与えられたルール群を基に目標とするルール群を生成するが，対象数が少ない場合，初期に与えられるルール群がその結論部に対応する決定クラスの特徴をうまく捉えていない可能性が高く，それを基にして抽出されるルール群は分類精度が低くなると考えられる．サンプリングの割合が 10%以上では，制限法，緩和法ともに従来法と同程度かそれ以上の分類精度を示している．特に，20%と 30%では緩和法の分類精度が四つの手法の中で最も良くなっている．

優先順位法について述べると，すべてのサンプリングの割合について，従来法と同じかそれよりも大きくなっており，また，対象数が多い場合 (60%から 80%) であっても従来法と有意差がある．よって，優先順位法のアルゴリズムで用いたデータの峰または谷を考慮する条件の選択が有効であるといえる．

6. おわりに

本研究では，入れ子構造を仮定したデータから上側和集合と下側和集合を結論部にもつ決定ルールを抽出する問題に対して，結論部の含意関係を考慮した決定ルール抽出方

表 3: 分類精度

割合 (%)	制限法 (%)	緩和法 (%)	優先順位法 (%)	従来法 (%)
1	70.89* ± 6.13	68.93 ± 6.40	71.19 ± 6.09	71.08 ± 6.12
2	77.65 ± 4.03	76.04 ± 4.48	78.04* ± 4.21	77.97 ± 4.17
3	80.81 ± 3.13	79.93 ± 3.62	81.46* ± 3.05	81.36 ± 3.10
4	83.08 ± 2.63	82.48 ± 3.06	83.51* ± 2.52	83.50 ± 2.55
5	84.49 ± 2.47	84.27 ± 2.66	85.01* ± 2.33	84.90 ± 2.33
6	85.63 ± 2.29	85.60 ± 2.49	86.12 ± 2.10	85.96 ± 2.17
7	86.60 ± 2.02	86.65 ± 2.22	86.98 ± 1.95	86.85 ± 1.95
8	87.41 ± 2.03	87.65* ± 2.01	87.83 ± 1.82	87.60 ± 1.85
9	88.29 ± 1.83	88.49* ± 1.88	88.59 ± 1.71	88.39 ± 1.73
10	88.92* ± 1.68	89.10 ± 1.74	89.12 ± 1.63	88.95 ± 1.58
20	92.91 ± 1.17	93.09 ± 1.15	92.89 ± 1.17	92.82 ± 1.15
30	95.01 ± 0.95	95.10 ± 0.88	95.00 ± 0.90	94.91 ± 0.93
40	96.26 ± 0.71	96.30 ± 0.70	96.28 ± 0.71	96.20 ± 0.72
50	97.17 ± 0.61	97.16 ± 0.61	97.17 ± 0.61	97.11 ± 0.61
60	97.73* ± 0.50	97.74* ± 0.50	97.78 ± 0.49	97.72 ± 0.50
70	98.18* ± 0.46	98.19* ± 0.43	98.22 ± 0.44	98.17 ± 0.45
80	98.53* ± 0.39	98.52* ± 0.38	98.55 ± 0.37	98.53 ± 0.40
90	98.78* ± 0.34	98.79* ± 0.35	98.80* ± 0.35	98.79 ± 0.35

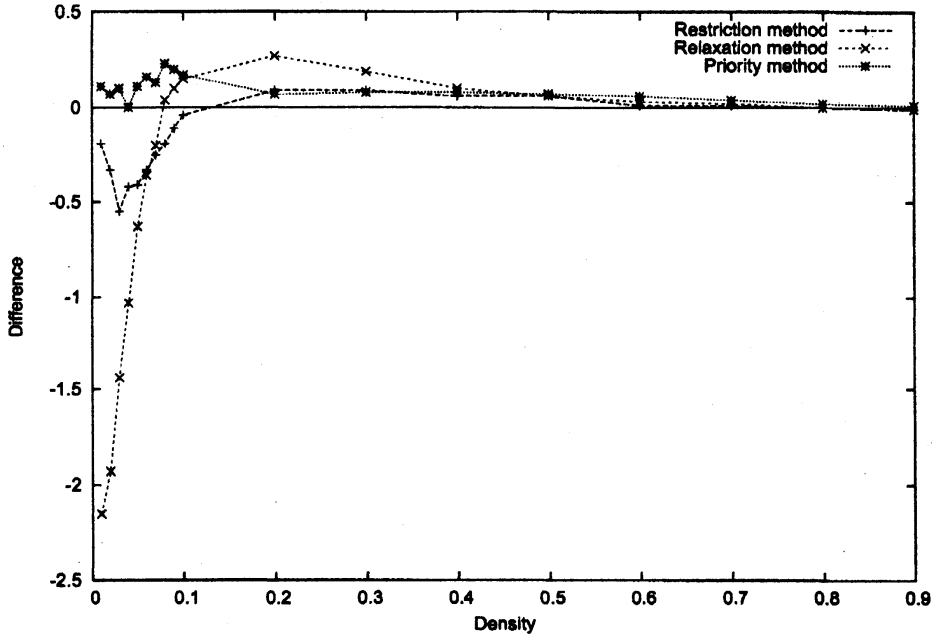

図 1: 各提案法と従来法の分類精度の平均値の差

法を三つ提案した．制限法と緩和法により，ある上側 (下側) 和集合に推論される対象が必ず下位 (上位) の上側 (下側) 和集合に推論されるルール群が抽出され，優先順位法はこれが完全に実現できないが従来法より高い割合でこれが成立する．また，抽出したルール群と下側和集合を結論部にもつルール群を用いた未知対象の分類方法も提案した．さらに，自動車の評価に関するデータを用いて数値実験を行った．結果として，優先順位法は，データ数に関係無く，従来法と同程度かそれ以上の分類精度を示し，制限法と緩和法は，極端に少ないデータでなければ，従来法の分類精度以上の結果を示した．今後の課題として，他のデータを用いた数値実験などがあげられる．

参考文献

[1] Asuncion, A and Newman, D. J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University California, School of Information and Computer Science.

[2] Błaszczyński, J., Greco, S. and Słowiński, R.: Multi-criteria classification - A new scheme for application of dominance-based decision rules, *European Journal of Operational Research*, Vol.181, Issue 3, pp.1030-1044, 2007

[3] Greco, S., Matarazzo, B. and Słowiński, R.: Rough Approximation by Dominance Relations, *International Journal of Intelligent Systems*, Vol.17, Issue 2, pp.153-171, 2002

[4] Grzymala-Busse, J. W.: LERS – A system for learning from examples based on rough sets, *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory* edited by R. Słowínski, Kluwer Academic Publishers, Dordrecht, pp.3-18, 1992.

[5] Grzymala-Busse, J. W. and Stefanowski, J.: Three Discretization Methods for Rule Induction, *International Journal of Intelligent Systems*, Vol.16, Issue 1, pp.29-38, 2001.

[6] Grzymala-Busse, J. W.: MLEM2 - Discretization During Rule Induction, *International Conference on Intelligent*, pp.499-508, 2003.

[7] 乾口 雅弘:ラフ集合による情報の解析, システム/制御/情報, Vol.49, No.5, pp.165-172, 2005.

[8] Pawlak, Z.: Rough sets, *Internat. J. of Inform. & Comput. Sci.*, Vol.11, No.5, pp.341-356, 1982.

[9] Pawlak, Z. and Skowron, A.: Rudiments of rough sets, *Information Sciences*, 177, Issue 1, pp.3-27, 2007.